Int J Life Cycle Assess (2011) 16:739–747
DOI 10.1007/s11367-011-0281-y

Simplified fate modelling in respect to ecotoxicological and human toxicological characterisation of emissions of chemical compounds

Morten Birkved · Reinout Heijungs

Received: 31 December 2010 / Accepted: 15 March 2011 / Published online: 19 April 2011

Abstract
Purpose The impact assessment of chemical compounds in Life Cycle Impact Assessment (LCIA) and Environmental Risk Assessment (ERA) requires a vast amount of data on the properties of the chemical compounds being assessed. The purpose of the present study is to explore statistical options for reduction of the data demand associated with characterisation of chemical emissions in LCIA and ERA.
Materials and methods Based on a USEtox™ characterisation factor set consisting of 3,073 data records, multi-dimensional bilinear models for emission compartment specific fate characterisation of chemical emissions were derived by application of Partial Least Squares Regression. Two sets of meta-models were derived having 63% and 75% of the minimum data demand of the full USEtox™ characterisation model. The meta-models were derived by grouping the dependent variables, the fate factors obtained from the USEtox™ data set and then selecting the independent chemical input parameters from the minimum data set, needed for characterisation in USEtox™, according to general availability, importance and relevance for fate factor prediction.
Results and discussion Each approach (63% and 75% of the minimum data set needed for characterisation in USEtox™) yielded 66 meta-models. In general, good correlation was obtained between the observed fate factors (those fate factors included in the USEtox™ data set) and the predicted fate factors (those fate factors obtained by the meta-models), and the validation regression coefficients were all in the range (R^2=0.41–0.96). The lower end of the regression coefficient range represents those few emission scenarios were the selected independent variables did not contain appropriate information. Hence, most meta-models yielded fate factors in good correlation with the observed fate factors and yielded correlation coefficients in the higher end of the range during validation. In general, the more data-demanding approach yielded the largest regression coefficients.
Conclusions The applied statistical approach illustrates that it is possible to derive meta-models from full fate and exposure models and that it is also possible to tailor the data demand of these meta-models according to various data and emission preferences. The results obtained in the study reveal that not all emission scenarios included in USEtox™ are exploiting the minimum data set equally and the minimum data set may thus in many cases contain underused data.

Keywords Approximated fate modelling · Fate modelling · Model approximation · Simplified characterisation · Simplified fate modelling · Simplified impact assessment · Underused fate parameters · USEtox

Responsible editor: Berlan Rodriguez-Perez

Electronic supplementary material The online version of this article (doi:10.1007/s11367-011-0281-y) contains supplementary material, which is available to authorized users.

M. Birkved (✉)
Department of Management Engineering,
Technical University of Denmark,
Lyngby 2800, Denmark
e-mail: birk@man.dtu.dk

R. Heijungs
Institute of Environmental Sciences, Leiden University,
Leiden 2300 RA, The Netherlands

1 Introduction

Assessment of toxic releases in environmental risk assessment (ERA) and Life Cycle Impact Assessment (LCIA)

often proceeds with multimedia fate and exposure models attached to models of dose–response relationships. The applicability of such models is, however obstructed by the fact that the environmental processes included in these models are complicated non-linear functions of a large number of parameters. Some of these are environmental parameters (such as the soil composition and the temperature), and other parameters are substance-specific (like the atmospheric degradation rate and the octanol–water partitioning coefficient). Especially substance-specific data are often hard to get, in particular for any of the thousands of lesser known and lesser common chemicals. This problem shows up in a marked way in LCIA, where the releases of hundreds or even thousands of toxic chemicals throughout a product life cycle (from mining to final disposal) are aggregated into a few impact categories, such as human toxicity and aquatic ecotoxicity. It is therefore attractive to seek ways to circumvent the use of these models, whilst keeping a close correspondence with the results of such models. In fact, it would be appealing if one could construct a model of these models, i.e. a simplified representation of the true model structure. In this paper, we will refer to this "simplified model of a model" as a "meta-model", and use the term "model" for the original thing which is supposed to be reflected in the meta-model.

In normal model construction, a model is supposed to be inspired by and validated against experimental results. In the meta-model construction that we are seeking to describe, validation should take place against the results of the original model. In other words, we suppose that the original model is available and that it has been validated appropriately for a suitable number of situations. The purpose of the meta-model is then to omit the use of the original model, and to use the meta-model whenever the data requirements of the original model cannot be met.

In ERA, an important class of meta-models is quantitative structure–activity relationships (QSARs) (see, e.g. Jensen 2006). QSARs are available to predict or estimate quantitative properties of chemicals on the basis of other quantitative properties of that chemical. In QSAR theory, the basic assumption is that many substance-specific parameters are correlated, at least for specific groups of substances. Estimation of missing parameters can therefore in principle take place on the basis of established QSARs. Although a data-demanding multimedia fate and exposure model might be run on the basis of a large number of QSAR estimates, the fact that so many estimates are combined into one overall model result leads to adopt a different approach. Our approach in this paper is based on the fate factors obtained from a newer and acknowledged multimedia fate and exposure model, in this case USEtox™ (Rosenbaum et al. 2008), and these results are statistically connected with a subset of selected input parameters that are needed for the USEtox™ model. The exact choice of the set of input parameters is made by scanning databases for availability of the individual parameters, combined with statistical information. The meta-model relationships themselves are established on a purely statistical basis. The normal recommendations (Cronin and Schultz (2003)) for calibration and validation of QSAR theory are followed in the approach presented.

2 Methods

The development of the meta-model proceeds through the following steps:

1. Grouping of dependent variables that can be modelled together
2. Division of the data set in calibration and validation set
3. Selection of data transformation and scaling
4. Calibration of the model
5. Location of appropriate X variables
   (a) Optimisation/trimming of model by deselection of insignificant X variables, recalibration and calculation of linear calibration coefficients
6. Validation of meta-model

The exact procedure on how to interpret the results from the individual steps is presented by Wold et al. (2001).

This section discusses the following elements in more detail:

- The multimedia fate and exposure model that served as a basis for the derivation of the meta-model (i.e. USEtox™)
- The data set that was used to create a calibration and validation set
- The statistical techniques that were used to derive the meta-model

2.1 The multimedia fate and exposure model

A central element in LCIA is the characterisation step (ISO 2006), where the calculation of the category indicator results takes place, in most cases on the basis of characterisation factors. As a general rule, the emitted amount of a certain chemical is multiplied by the characterisation factor that connects that chemical to a certain impact category, after which an aggregation across chemicals within one impact category is performed:

$$\mathrm{CIR}_j = \sum_j \mathrm{CF}_{ji} m_i \qquad (1)$$

where m_i is the amount of a certain chemical released to compartment i, CF_{ji} is the characterisation factor that

connects the chemical in compartment i to impact category (and thus the compartment in/by which the effect is manifested) j, and CIR_j is the category indicator result for impact category j. The characterisation factors encapsulate the information on the fate, exposure and effect of the chemical; it is derived from the characterisation model.

In the toxicity-related impact categories, the characterisation factor is built with a number of separate elements (see also Rosenbaum et al. 2008):

- The aspect of fate, symbolised by the fate factor (FF);
- The aspect of exposure or intake, symbolised by the exposure factor (XF);
- The aspect of effect, symbolised by the effect factor (EF); and
- The combined aspect of fate and human exposure, symbolized the intake fraction (iF).

The aspect of exposure covers several intake routes for humans, such as inhalation of air and ingestion through crops, fish and dairy. For ecotoxicity, the aspect of intake is left out, and the impacts are defined right in the receiving compartment itself. In this way, two structures of the characterisation factor can be discerned:

$$CF_{ji} = EF_j XF_J FF_{ji} \quad (2)$$

for ecosystems, with j = freshwater and i = emission compartment for the impact category freshwater aquatic ecotoxicity, j = soil for the impact category terrestrial ecotoxicity, etc., and

$$CF_{ji} = \sum_l \sum_k EF_{jl} XF_{lk} FF_{ki} = \sum_l \sum_k EF_{jl} iF_{li} \quad (3)$$

for man, where j = man for the impact category human toxicity, and where k represents receiving compartments (air, freshwater, soil, etc.) and l represents intake routes (air, crops, fish, dairy, etc.).

As is evident from both characterisation approaches, the fate factor plays a central role in the characterisation of both ecotoxicological and human toxicological impact potentials.

The effect factors are determined from measured or estimated (usually by QSAR) effect measures, the fate factors all have to be calculated by multimedia fate modelling. The way the fate factors are modelled thus determines the fate-related data demand of a chemical characterisation in LCA and thus plays a crucial role in the data demand related to characterisation of chemical emissions in LCA. The exposure factors could be approached similar to the fate factors, but in this study, we focus solely on simplification of the fate factor modelling. To obtain a complete simplified methodology, similar simplified meta-models for the exposure factors are needed. Effect factors on the other hand can be obtained via alternative existing sources like US EPA (2009).

2.2 The data set

As part of the USEtox™ documentation, Huijbregts et al. (2010) has made available the substance-specific data for 3,073 organic chemicals along with the model results including fate factors. This data set served as a basis for the statistical model derivations for this paper.

In general, the fate factors on organic chemicals depend on a large number of physico-chemical properties, such as molecular weight, water solubility, octanol–water partition coefficient and compartment specific degradation rates. For the physico-chemical data, some parameters are widely available (such as the molecular weight and water solubility) while other parameters (such as compartment specific degradation rates) are available for only a few substances. Generally and not surprisingly, the pattern is that the cheaper it is to measure a property, the more likely it is that the property can be located for a given chemical.

All of the 3,073 organic chemicals in the above described dataset were selected for inclusion in our study and the fate factors on all 3,073 chemicals for emissions to continental urban air, continental rural air, continental freshwater, continental seawater, continental natural soil and continental agricultural soil were compiled, in total six emission compartments. The USEtox™ compound specific fate factors consist of 66 parameters for each chemical (11 fate factors for each emission compartment), of which all are needed to perform a full characterisation (i.e. calculation of characterisation factors for emission to all possible emission compartments and final compartments) in USEtox™. The descriptive statistics of the dependent and independent variables used as basis for the development of the meta-model are presented in Tables 1 and 2 in Appendix A.

As the grouping of the dependent variables can be done in various ways, we decided to let data availability determine the grouping of the variables and thereby optimize the applicability of the meta-model data wise. In this way, two meta-model derivation approaches were isolated. Basically, we decided to derive 66 meta-models (one for each final compartment) grouped according to emission/receiving compartment (i.e. six meta-model groups) based on five or six (approach 1 and 2) of the eight parameters included in the USEtox™ minimum dataset, equal to app. 63% and 75% of the original model minimum data demand. The six selected parameters were molecular weight, octanol–water partition coefficient, vapour pressure at 25°C, water solubility at 25°C, degradation rate in air and/or (approach 1/2) degradation rate in water. In the first approach which relies on five parameters, the degradation rate of the two selected degradation rates to be included was determined based on meta-model group importance of the two degradation rates, in such a way that only the degradation rate having the highest model importance was included.

Since linear and bilinear models assume linear data relationships, and since the variables selected for the model derivation are of physical, chemical and/or biological origin, and hence have a tendency to display skewed distributions, the variables have to be transformed in such a way that dependent variables can be assumed to be a linear function of the independent variables. Fate factors can easily span ten orders of magnitude (see Table 2 in Appendix A) or more due to the large differences in properties of the chemicals, and such large differences in scale can create a biased picture in the least squares fit of regression analysis. To raise the probability of linear dependency between the dependent and independent variables, these were logarithmically transformed. We used the ^{10}Log, abbreviated as Log. Using logarithmic transformation further reduces the scale differences to a much narrower range, with a much lower number of possibly influential data points. An additional advantage of the logarithmic transformation is that the differences in scale of the fate factors are reduced. After all, one should realise that fate factors typically are specified per kilogram emission of chemical compound.

The results of the statistical estimation techniques (see below) are scale dependent, i.e. they depend on the magnitude of the variables. To avoid giving certain independent parameters, more weight than other due to their magnitude, all parameters are scaled to unit variance, thereby making the results independent of the units of the variables. All parameters are transformed to unit variance by division with the standard deviation around the mean of the variables.

2.3 The statistical techniques

The statistical technique used to derive the meta-models is related to two different activities:

- Estimation of the most appropriate model and its coefficients through regression analysis; and
- Validation of the predictive capability of the model by methods taken from QSAR-practice.

Below, both aspects are discussed.

The derivation of the meta-model from the model results takes place on the basis of regression analysis. This widely used statistical tool for identifying multilinear relationships between several independent variables (say, x_1. x_2, …) and one dependent variable (say, y) is assumed to be known to the reader. The general form of such a model is:

$$y = a_o + a_1x_1 + a_2x_2 + \ldots \quad (4)$$

where $a0$, $a1$, $a2$, … are constants that are to be estimated. The nature of the problem dealt with in this paper forces us to reconsider one assumption of the ordinary least-squares regression model: the independence of the independent variables. In the present context, the set of independent variables is formed by parameters such as the molecular weight, the vapour pressure, the solubility, the octanol–water partition coefficient and the degradation rate in water. As is known from QSAR theory, there are approximate relationships between some of these variables. Hence, the assumption of independence in an actual data set will be violated to some extent.

Alternatives to the classical regression model have been developed, partly with the purpose of obviating the assumption of independence. These variants are known under names as principal component regression (see Esbensen 2000; Vigneau et al. 1997), ridge regression (see Vigneau et al. 1997) and partial least-squares regression (PLSR) (see Esbensen 2000). In the sequel of this paper, the PLSR method has been employed for the following additional reasons:

- It allows for the modelling of dependent, noisy variables (even with missing observations)
- It allows for the modelling of multiple dependent variables (say, y_1, y_2, …)
- It is available in software (e.g. SIMCA-P+ Umetrics (2009)) that allows for the identification of the optimal set of independent variables, as well as the optimal transformations of these variables

Please refer to Appendix B for a short introduction to PLSR theory. For a more detailed introduction to the PLSR theory, please refer to Martens and Dardenne (1998), Martens and Martens (2001), Esbensen (2000) and Wold et al. (2001).

Apart from selecting the best model and calibrating the coefficients of the model, one should determine the quality of the predictive capability of the model. Regression models of the type derived in this paper are comparable to the models typically encountered in QSAR. In this study, we calculated R^2, the coefficient of determination, which can be interpreted as the fraction of variance that is explained by the model, Q^2 (cum) which is the cumulative fraction of the total variation of x or y that can be predicted by the components (estimated by cross-validation) and the root mean square error of prediction, which is the standard deviation of the predicted residuals or more precisely errors and is calculated as presented in Appendix C.

Common to all QSAR techniques are the non-standardised recommendations for the development and use of QSARs. All of these recommendations are however aimed at QSARs developed to estimate structural–biological and/or structural–chemical activity relationships such as biodegradation, toxicity and physical chemical properties. What the models

presented in this paper are designed to do is to estimate fate factors to facilitate the calculation of characterisation factors in LCA. In this way, the meta-models do not model structure–activity relations but a chemical property–model relation. In general, there are no formal guidelines for developing QSARs (Cronin and Schultz 2003) but politically determined recommendations on common QSARs have been presented (see, e.g. European Chemical Bureau 2003). It is unclear whether recommendations like these apply to the type of models presented here.

A crucial point in QSAR and general model development is the approach taken for validating the model. Based on the aspects presented in Appendix C, it was concluded that test set validation would give the best estimate of the prediction performance of models of the type developed here based on the data set summarised in Tables 1 and 2 in Appendix A. In our study, 616 records (20%) of the data set (3,073 records in total) were randomly selected and used as validation set.

3 Results

As presented in Tables 1 and 2, 2×66 emission and effect compartment specific meta-models were derived applying both the 63% and 75% data demand approach. All 112 validation plots for all possible combinations of included emission compartments and effect compartments are presented in Appendices D (approach 1–63% data demand) and E (approach 2–75% data demand).

Based on the meta-model coefficients obtained from both approaches (see Appendix F and G), it is possible to construct the approximated multiple linear models for all combinations of emission and effect compartment. We, do however recommend to use the exact meta-models (please refer to Appendix H for the complete SIMCA-P + files). Below, example illustrates how the approximated factor for emission and effect in urban air is constructed according to the approach 1 results (see Table 1 in Appendix F).

$$\begin{aligned}\mathrm{Log}\left(\mathrm{FF}_{\mathrm{airU,airU}}\right) = &-2.42 - 6.55\times10^{-1}\times\mathrm{Log(Mw)} + 1.19\times10^{-2}\times\mathrm{Log(Kow)} + 1.29\times10^{-4}\times\mathrm{Log(Pvap25)}\\ &+7.79\times10^{-3}\times\mathrm{Log(Sol25)} - 5.58\times10^{-1}\times\mathrm{Log(kdegA)} + 1.68\times10^{-1}\times\mathrm{Log(Mw)}^2\\ &-2.19\times10^{-4}\times\mathrm{Log(Kow)}^2 + 1.73\times10^{-4}\times\mathrm{Log(Pvap25)}^2 + 8.60\times10^{-5}\times\mathrm{Log(Sol25)}^2\\ &-5.83\times10^{-2}\times\mathrm{Log(kdegA)}^2 + 4.76\times10^{-3}\times\mathrm{Log(Mw)}\times\mathrm{Log(Kow)} - 7.26\times10^{-3}\mathrm{Log(Mw)}\times\mathrm{Log(Pvap25)}\\ &+8.38\times10^{-4}\times\mathrm{Log(Mw)}\times\mathrm{Log(Sol25)} - 2.57\times10^{-2}\times\mathrm{Log(Mw)}\times\mathrm{Log(kdegA)}\\ &+1.72\times10^{-4}\times\mathrm{Log(Kow)}\times\mathrm{Log(Pvap25)} - 7.66\times10^{-5}\times\mathrm{Log(Kow)}\times\mathrm{Log(Sol25)}\\ &+3.63\times10^{-3}\times\mathrm{Log(Kow)}\times\mathrm{Log(kdegA)} - 5.01\times10^{-4}\times\mathrm{Log(Pvap25)}\times\mathrm{Log(Sol25)}\\ &-2.84\times10^{-3}\times\mathrm{Log(Pvap25)}\times\mathrm{Log(kdegA)} + 2.02\times10^{-3}\times\mathrm{Log(Sol25)}\times\mathrm{Log(kdegA)}\\ &-1.51\times10^{-2}\times\mathrm{Log(Mw)}^3 + 9.66\times10^{-6}\times\mathrm{Log(Kow)}^3 + 6.74\times10^{-6}\times\mathrm{Log(Pvap25)}^3\\ &+3.96\times10^{-6}\times\mathrm{Log(Sol25)}^3 - 2.07\times10^{-3}\times\mathrm{Log(kdegA)}^3\end{aligned} \quad (5)$$

Where: FF = fate factor, airU = urban air (s/m^3), Mw = molecular weight (g/mol), Kow = octanol–water partition coefficient (unit less), Pvap25 = vapour pressure at 25°C (Pa), Sol25 = water solubility at 25°C (mg/L), kdegA = degradation rate in air (1/s), kdegW = degradation rate in water (1/s). Notice that all numbers representing the coefficients a_0, a_1, etc. have a dimension. Figure 1 shows an example validation plot of the observed and the predicted values of the fate factor for emission and effect in urban air with approach 1.

4 Discussion

The summarised regression statistics on the 2×66 derived fate factor meta-models presented in Tables 1 and 2, indicates that a large amount of the observed variance in the fate factor can be explained by the PLSR derived linear models. As presented in Tables 1 and 2, it seems plausible to assume that the most important fate pathway is the compartment-specific degradation process in the emission compartment (e.g. degradation in air by emission to air).

In general, it is possible to explain a large part of the variance in the fate factors of the 11 effect compartments modelled in USEtox™ by the PLSR-derived linear models summarised in Tables 1 and 2. In the air emission scenarios (emission to continental urban or continental rural air), approach 1 achieves reasonable regression coefficients (R^2=0.61–0.86) while the slightly more data-demanding (app. data demand increase = 12%) approach 2 yields reasonable, however noticeable improved regression coefficients (R^2=0.71–0.87). Highest observed correlation between observed and predicted fate factors by emission to air compartments are observed for air effect compartments (exposure via urban air, continental air and global air). In the water emission scenarios (emission to continental fresh water and continental sea water), approach 1 achieves good regression coefficients (R^2=0.62–0.93) while the slightly more data-demanding approach 2 yields even better regression coefficients (R^2=0.78–0.95). Highest observed

Table 1 Validation results of the validation of the 66 meta-models obtained according to model approach 1 (making use of 63% of the minimum data set in USEtox™) grouped according to emission compartment

Final compartment	Emission compartment											
	Emission to continental urban air		Emission to continental rural air		Emission to continental freshwater		Emission to continental seawater		Emission to continental natural soil		Emission to continental agricultural soil	
	R^2	RMSEP	R^2	RMSEP	R^2	RMSEP	R^2	RMSEP	R^2	RMSEP	R^2	RMSEP
Urban air	0.72	1.3×10^{-5}	0.85	6.02×10^{-9}	0.93	8.47×10^{-9}	0.93	5.81×10^{-10}	0.94	9.29×10^{-9}	0.94	9.47×10^{-9}
Continental air	0.86	1.6×10^{2}	0.86	1.80×10^{2}	0.93	2.25×10^{2}	0.93	1.70×10^{1}	0.94	2.92×10^{2}	0.79	2.93×10^{2}
Continental freshwater	0.67	1.1×10^{2}	0.68	1.61×10^{1}	0.62	1.21×10^{2}	0.88	3.92×10^{-1}	0.53	1.57×10^{1}	0.53	1.58×10^{1}
Continental seawater	0.61	1.3×10^{4}	0.62	3.51×10^{3}	0.81	5.76×10^{1}	0.89	1.15×10^{2}	0.41	7.72	0.41	7.73
Continental natural soil	0.66	1.8×10^{9}	0.68	6.60×10^{8}	0.89	2.28	0.89	1.96	0.62	1.48×10^{4}	0.71	5.78
Continental agricultural soil	0.66	1.8×10^{9}	0.68	6.60×10^{8}	0.89	2.28	0.89	1.96	0.71	6.06	0.62	1.50×10^{4}
Global air	0.86	2.7×10^{5}	0.85	3.70×10^{5}	0.89	4.45×10^{5}	0.91	3.36×10^{4}	0.94	5.44×10^{5}	0.80	5.49×10^{5}
Global freshwater	0.76	5.2×10^{-2}	0.76	4.56×10^{-2}	0.85	4.30×10^{-4}	0.86	1.03×10^{-2}	0.89	4.43×10^{-4}	0.89	3.09×10^{-4}
Global oceanic seawater	0.71	6.6×10^{4}	0.71	1.72×10^{4}	0.72	1.84×10^{2}	0.89	9.19×10^{1}	0.42	5.33×10^{1}	0.42	5.08×10^{1}
Global natural soil	0.74	1.8×10^{6}	0.75	6.10×10^{5}	0.86	7.99×10^{-2}	0.88	1.25×10^{-1}	0.72	2.44	0.72	2.16
Global agricultural soil	0.74	1.8×10^{6}	0.75	6.10×10^{5}	0.86	7.99×10^{-2}	0.88	1.25×10^{-1}	0.72	2.44	0.72	2.16
Average R^2	0.72	–	0.74	–	0.84	–	0.89	–	0.71	–	0.69	–
Q^2(cum)	0.72		0.73		0.84		0.89		0.77		0.77	
Variables applied	Mw		Mw		Mw		Mw		Mw		Mw	
	Kow		Kow		Kow		Kow		Kow		Kow	
	PVap25		PVap25		PVap25		PVap25		PVap25		PVap25	
	Sol25		Sol25		Sol25		Sol25		Sol25		Sol25	
	kDegA		kDegA		kDegW		kDegW		kDegA		kDegA	

R^2 coefficient of determination, *RMSEP* root mean square error of prediction, Q^2 (cum) cumulative fraction of the total variation of x or y that can be predicted by components, *Mw* molecular weight, *Kow* octanol–water partition coefficient, *PVap25* vapour pressure at 25°C, *Sol25* water solubility at 25°C, *kDegA* degradation rate in air

Table 2 Validation results of the validation of the 66 meta-models obtained according to model approach 2 (making use of 75% of the minimum data set in USEtox™) grouped according to emission compartment

Final compartment	Emission compartment											
	Emission to continental urban air		Emission to continental rural air		Emission to continental fresh water		Emission to continental sea water		Emission to continental natural soil		Emission to continental agricultural soil	
	R^2	RMSEP	R^2	RMSEP	R^2	RMSEP	R^2	RMSEP	R^2	RMSEP	R^2	RMSEP
Urban air	0.76	1.1×10^{-5}	0.80	9.24×10^{-9}	0.95	8.93×10^{-8}	0.95	1.28×10^{-9}	0.91	6.60×10^{-6}	0.91	6.60×10^{-6}
Continental air	0.87	2.0×10^{2}	0.79	2.74×10^{2}	0.95	1.26×10^{3}	0.96	3.36×10^{1}	0.77	5.31×10^{4}	0.77	5.31×10^{4}
Continental freshwater	0.71	4.4×10^{1}	0.69	2.13×10^{1}	0.60	1.37×10^{2}	0.90	3.70×10^{-2}	0.57	2.34×10^{1}	0.57	2.34×10^{1}
Continental seawater	0.71	4.5×10^{3}	0.68	3.28×10^{3}	0.78	7.93×10^{1}	0.90	9.50×10^{1}	0.56	7.14	0.56	7.14
Continental natural soil	0.78	2.2×10^{7}	0.73	2.30×10^{8}	0.91	1.51×10^{2}	0.92	8.04	0.75	1.15×10^{5}	0.69	7.93
Continental agricultural soil	0.78	2.2×10^{7}	0.73	2.30×10^{8}	0.91	1.51×10^{2}	0.92	8.04	0.69	7.93	0.75	1.15×10^{5}
Global air	0.86	4.4×10^{5}	0.78	5.94×10^{5}	0.80	4.92×10^{5}	0.95	3.40×10^{4}	0.79	5.09×10^{5}	0.79	5.09×10^{5}
Global freshwater	0.79	8.4×10^{-2}	0.76	8.86×10^{-2}	0.89	5.99×10^{1}	0.90	5.06×10^{-3}	0.87	4.06×10^{-1}	0.87	4.06×10^{-1}
Global oceanic seawater	0.79	1.7×10^{4}	0.77	1.46×10^{4}	0.80	8.96×10^{1}	0.91	8.41×10^{1}	0.61	7.77×10^{1}	0.61	7.77×10^{1}
Global natural soil	0.82	4.1×10^{4}	0.80	3.30×10^{5}	0.92	8.80	0.93	4.88	0.90	1.03	0.90	1.03
Global agricultural soil	0.82	4.1×10^{4}	0.80	3.30×10^{5}	0.92	8.80	0.93	4.88	0.90	1.03	0.90	1.03
Average R^2	0.79		0.76		0.86		0.92		0.76		0.76	
Q^2 (cum)	0.76		0.73		0.86		0.91		0.80		0.80	
Variables applied	Mw		Mw		Mw		Mw		Mw		Mw	
	Kow		Kow		Kow		Kow		Kow		Kow	
	PVap25		PVap25		PVap25		PVap25		PVap25		PVap25	
	Sol25		Sol25		Sol25		Sol25		Sol25		Sol25	
	kDegA		kDegA		kDegA		kDegA		kDegA		kDegA	
	kDegW		kDegW		kDegW		kDegW		kDegW		kDegW	

R^2 coefficient of determination, *RMSEP* root mean square error of prediction, Q^2 (cum) cumulative fraction of the total variation of x or y that can be predicted by components, *Mw* molecular weight, *Kow* octanol–water partition coefficient, *PVap25* vapour pressure at 25°C, *Sol25* water solubility at 25°C, *kDegA* degradation rate in air, *kDegW* degradation rate in water

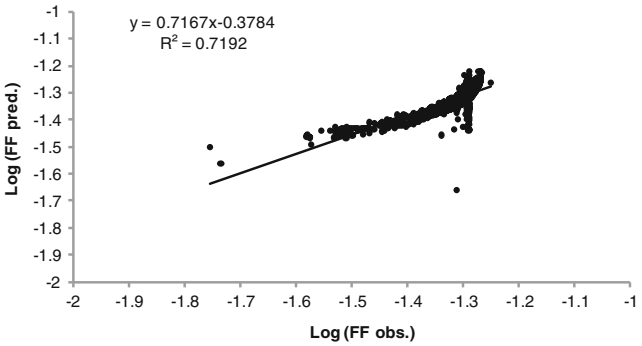

Fig. 1 Validation plot (n=616) for the meta-model capable of predicting fate factors for characterization of effects in urban air by emission to continental seawater. This model was obtained by application of model derivation approach 1, applying 63% of the minimum data set for USEtox™. *FFobs.* observed fate factors [obtained from the USEtox™ data set (Huijbregts et al. 2010)] and *FFpred.* predicted fate factors (obtained by application of the meta-model)

correlation between observed and predicted fate factors by emission to water compartments are observed for air effect compartments (exposure via urban air, continental air and global air). We conjecture that this reflects the much simpler role of the air compartments in fate models. In the soil emission scenarios (emission to continental natural soil and continental agricultural soil), approach 1 achieves poor to good regression coefficients (R^2=0.41–0.94) while the slightly more data-demanding approach 2 yields improved regression coefficients in the range (R^2=0.56–0.90). Highest observed correlation between observed and predicted fate factors by emission to soil compartments are again observed for air effect compartments (urban air, continental air and global air), probably for the same reason as conjectured above.

The purpose of the linear models is to mimic the way USEtox™ treats certain combinations of independent variables, and whether the linear models have been derived from fate factors calculated from real independent data (measured compounds specific input) or from estimated data sets will most likely not influence the parameterisation of the derived linear meta-models significantly. The lack of importance of data origin is caused by the fact that USEtox™ as any other model treats estimated and measured data sets the same way. In the present case, we did not use soil and sediment degradation rates, both parameters are complicated to estimate precisely and even more importantly, measured degradation rates for these two compartments are belonging to the more exclusive group of laboratory fate data, only available on chemicals with known problematic properties and used in considerable quantities.

The presented meta-models have all been derived from data on a broad group of organic chemicals. It may well be that deriving a separate meta-model for, say, chlorinated hydrocarbons, yields an even better model performance. This applies in particular to metals and other speciating inorganic chemicals, which are in many respects different from organic chemicals.

5 Conclusions

As presented, it has been demonstrated that it is possible to explain large amounts of the variance observed in fate factors obtained from USEtox™ characterisation by simple linear models of the same type as presented in Eq. 5. Deriving simple models from complex models potentially opens for the creation of model compilations, suitable for specific data availability combinations or characterisation of specific emission situations. Statistical derivation assures a significant correspondence between full and meta-models ensuring compatibility and aggregation potential of the results.

An important aspect of deriving the presented type of meta-models is that the relevance of the individual independent data is illuminated. The full SIMPCA-P + 12.0 model (app. 50 Mb) is available from the corresponding author upon request.

Acknowledgements The authors would like to thank Arjan de Koning, Institute of Environmental Sciences, Leiden University and Professor Michael Z. Hauschild, Department of Management Engineering Technical University of Denmark for their support in the initial stages of the study.

References

Cronin MTD, Schultz TW (2003) Pitfalls in QSAR. J Mol Struct Theochem 622(1–2):39–51

Esbensen KH (2000) Multivariate data analysis—in practice. Camo ASA, Oslo

European Chemical Bureau (2003) Technical guidance documents in support of Commission directive 96/67/EEC on risk assessment for new notified substances, Commission directive (EC) No. 1488/94 on risk assessment for existing substances, Directive 98/8/EC of the European Parliament and of the Council concerning the placing of biocidal products on the market. Part III. European Commission, Joint Research Centre, Ispra, http://publications.jrc.ec.europa.eu/repository/bitstream/111111111/1212/1/EUR%2019909%20EN.pdf

Huijbregts MAJ, Margni M, Van de Meent D et al. (2010) USEtox™ Chemical-specific database: organics. Report published by the USEtoxTM team. Available upon request at http://www.usetox.org/

ISO (2006) Environmental management—Life cycle assessment—Requirements and guidelines (ISO 14044). International Organization for Standardization, Geneva

Jensen F (2006) Introduction to computational chemistry. Wiley, New York

Martens H, Dardenne P (1998) Validation of regression in small data sets. Chemom Intell Lab Syst 44:99–121

Martens H, Martens M (2001) Multivariate analysis of quality: an introduction. Wiley, New York

OECD (2007) Guidance on the validation of (quantitative) structure-activity relationship [(Q)SAR] models. Organisation for Eco-

nomic Co-operation and Development, Paris, http://www.oecd.org/officialdocuments/displaydocumentpdf?cote=env/jm/mono(2007)2&doclanguage=en

Rosenbaum RK, Bachmann TM, Gold LS, Huijbregts MAJ, Jolliet O, Juraske R, Koehler A, Larsen HF, Macleod M, Margni M, McKone TE, Payet J, Schumacher M, Van de Meent D, Hauschild MZ (2008) USEtox –the UNEP-SETAC toxicity model: recommended characterisation factors for human toxicity and freshwater Ecotoxicity in LCIA. Int J Life Cycle Assess 13:532–546

Umetrics (2009) SIMPCA-P+, version 12.0.1.0. Umetrics AB, Umeå

US EPA (2009) Estimation Program Interface Suite (EPI Suite), version 4.00. United States Environmental Protection Agency, Washington DC

Wold S, Sjöström M, Eriksson L (2001) PLS-regression: a basic tool for chemometrics. Chemom Intell Lab Syst 58:109–130

Vigneau E, Devaux MF, Qannari EM, Robert P (1997) Principal component regression, ridge regression and ridge principal component regression in spectroscopy calibration. J Chemometr 11:239–249